I Know Your Family: A Hybrid Information Retrieval Approach to Extract Family Information from Microblogs

Jamuna Gopal

Submitted to the graduate degree program in Electrical Engineering & Computer Science and the Graduate Faculty of the University of Kansas School of Engineering in partial fulfillment of the requirements for the degree of Master of Science

Thesis Committee:

Dr. Bo Luo: Chairperson

Dr. Jerzy W. Grzymala-Busse

Dr. Prasad Kulkarni

Date Defended

The Thesis Committee for Jamuna Gopal certifies
that this is the approved version of the following thesis:

I Know Your Family: A Hybrid Information Retrieval Approach to Extract Family Information from Microblogs

Committee:

_____

Chairperson

_____

_____

_____

Date Approved

Abstract

With the growing popularity of social networks, large amounts of personal information have been made available over the Internet. The aim of this thesis is to identify the family related information of a person from their microblogs (Twitter). We use their personal details, tweets and their friends' details in order to achieve this. Since, we deal with the modern world short text data; we have used a hybrid information retrieval methodology taking into account the Parts of Speech of the data, Phrase Similarity and the Semantic Similarity of the data along with the openly available twitter data. The future use of this research is to develop a Client Side protection tool that will help users validate the data to be posted for privacy breech.

I like to dedicate this work to my mother.

Acknowledgements

First and foremost, I would like to thank my advisor in this project, Dr. Bo Luo, for his patient guidance and encouraging demeanor. He inspired me greatly. This project is a result of his vision. His kindness, availability, motivation and goal mindedness helped me remain consistent and focused. This research is dedicated to Dr. Bo Luo's positive thinking.

I would like to thank my committee members, Dr.Prasad Kulkarni and Dr. Jerzy W. Grzymala-Busse for being supportive. A special appreciation to Hariprasad Sampathkuma and Wenrong Zeng for the guidance provided. I would like to thank the University Of Kansas for its research program. The facilities have conveniently proven useful during each step of my research. My role as a Teaching Assistant not only helped me financially, but also sculpted me into being a more dynamic and responsible student.

Finally, I would like to express my gratitude to my family and friends, without whom I would have never dreamt of being a student of science. Their moral support and motivation at every stage has driven the consistency of purpose in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

Communication has evolved greatly with the recent Online Social Networking Sites (OSN). Online Communication and information exchange is setting the trend as opposed to face-to-face communication. As of September 2013, 73% of the internet users use social networking sites [1]. This has increased phenomenally over the years since it was just 5% of the users as of February 2005 [1]. With the increase in the users count, the data that is available publicly has increased numerous folds. This data includes personal, employment, education, relationship, family related information about the users. Figure 1.1, 1.2 below are 2 examples of data and communication over micro-blogs. Figure 1.1 is an example of a tweet message that a user broadcasts over to the public. Figure 1.2 is the About page in Facebook, where the details of the user can be found.

Figure 1.1. Twitter - Tweet Example

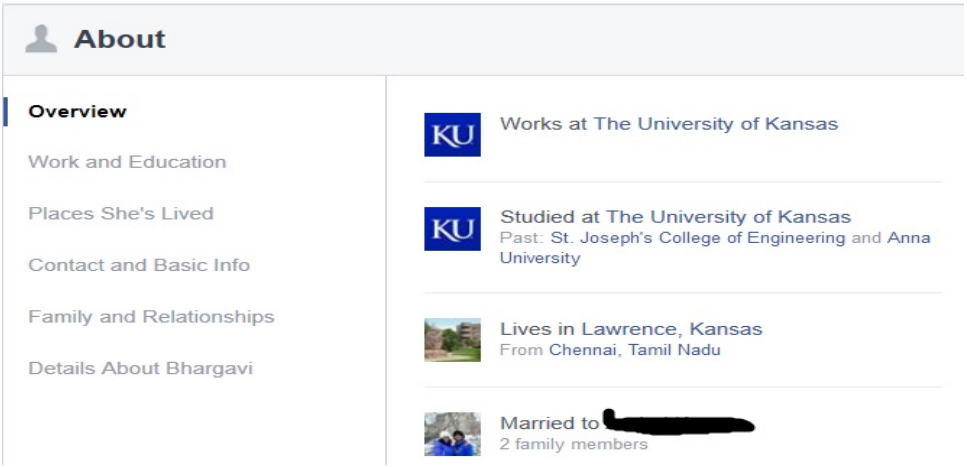

Figure 1.2. Facebook - About Page

This data is mined efficiently to handle user problems and to benefit users. For example, Facebook a popular OSN keeps track of the users birthday and sends notification to friends of the user. Online sites like JCPenny, Kohls, Jabong sends Gift vouchers to registered users on their special days like Birthday and Anniversary. Mining this public data also has commercial purpose. Some of the examples under this category are given below.

- A baby product retailer mines social data to choose appropriate audience for their online ads. For example, messages like "My babys first birthday", "Happy birthday niece. Cant believe its been a year already" will help the company to post ads related to baby products and baby gifts.

- A location restaurant mines the social data to find users in their locality to publish the new menu and offers.

- Banks and other financial businesses use the social data to build the trust score of a user for granting loan.

- Book, Audio, Movies, Television, every industry now depends on feedback

from this public data to improve their market.

Customized Ad's though liked by some users, poses a privacy threat to all users which many are not aware of. To enhance their market, online advertisement companies mine this publicly available data to predict user attributes such as age, gender, marital status, number of children, interests etc. They use these mining results to post enticing Ad's that suit the user's features on to the user's page. Although this is a good establishment, it is a privacy hole since private data is used for commercial gain.

Though these micro-blog mediums are greatly useful to express opinions and share interests, the public-centric data that they expose holds a major challenge to privacy. Tweets, Check-ins, and status about the users current travel location expose their home location to burglary. A message with birthday, anniversary wish exposes user's age and his/her family information to online stalkers. Messages containing user tag's exposes the tagged user along with their relationships. Cover Pictures and photo tag information reveals user's identity to the Internet world. Most users do not post their Address, SSN and phone numbers as public data [2]. A research by [3] explains how an SSN number could be identified with just users birthday and their place of birth information. This data helps to identify the first 5 digits of SSN, and hence only the last 4 digits is held private. A 4 digit code combination is easy to break which makes the social data a huge privacy constraint. Hence an online stalker with little hacking capability and ample time can figure out every detail about the user with the data available online.

The underlying idea behind this research is to address the problem of protecting family related information about the user. We have developed an algorithm that crawls through the user data to identify the family related information present

on the user feed. Although this is itself a privacy violation, this algorithm is developed with a good intention of being extended in the future to identify privacy threats in a message before the data goes public. Manual message identification on micro-blogging sites is a cumbersome process since less than 1% of the data publicly available is related to the family. Our focus to predict the family attributes is based on the publicly available Twitter data. Most of the data present on Twitter uses shortened words; hence mining using dictionary match is almost impossible. Numerous patterns are present and every tweet has a unique way of conveying its message. Hence our prediction algorithm considers multiple features about the tweet comprising of parts-of-speech tags, similarity with respect to the words, and similarity with respect to family relations. Thus, our prediction model is a binary classifier, which classifies each message as either "Family sensitive" or "Insensitive" based on the weighted sum of the individual features considered.

In an overview, our algorithm takes all the tweets of the user, user details, friends list, processes each tweet across multiple features and outputs the tweets that are related to the family.

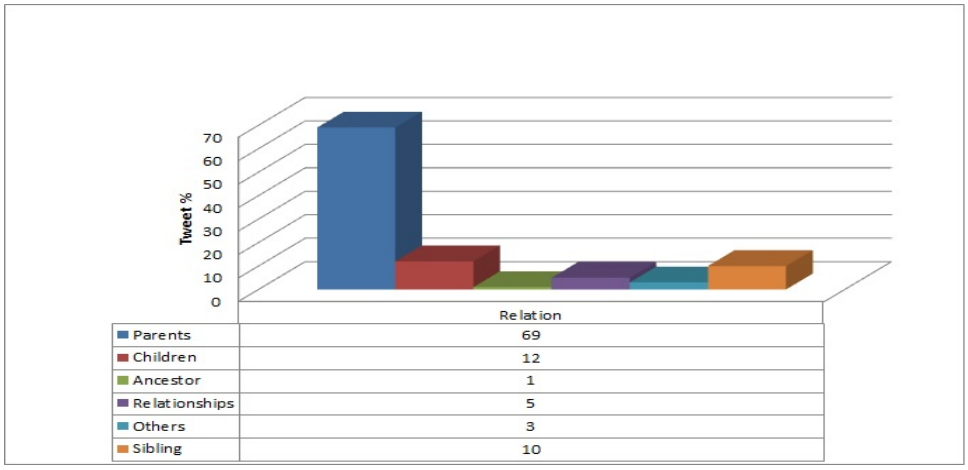

Figure 1.3. Family Data Distribution

The figure 1.3 above shows the general family related tweet trend based on our data-set. The following are our contributions,

- We are the first to attempt family related attribute prediction in Twitter data.
- The mechanism involves similarity measures and relationship analyses.Hence, the multi-feature mechanism is superior to the word-search algorithm.
- We observed the precision rate to be over 83% for identifying family tweets. Noise content reduction rate is over 62%.

The rest of the thesis is organized as follows. Introduction to micro-blogs and related works in this area are discussed in Chapter 2. The problem under study and data related challenges are discussed in Chapter 3. A detailed analysis of our algorithm and its sub-components is discussed in Chapter 4. Results and performance analyses are presented in Chapter 5. The evaluation of the models and future scope is considered in Chapter 6.

Chapter 2

Background and Related Work

2.1 About Microblog's and Twitter

Blogging is an age old methodology which is primarily used to express opinions and thoughts. Blogging is famous across authors, book reviewers and social writers since blogs do not have limitation on content size. Though blogging is still being used, Internet users have moved their focus to Micro-Blogging, where the content of the blog is a small word set of less than 50 words. Micro-blogging became a hit because of its less time consumption and broader audience reach. Micro-blogs - "allow users to exchange small elements of content such as short sentences, individual images, or video links" [4]. As the definition suggests, we are dealing with short text data which can be only up to certain fixed character length.

Every OSN has its own convention and terminology for providing communication. The Micro-blogging site that we primarily deal with in this research is Twitter. Twitter is a popular OSN which has registered 117 million active users [1]. All user accounts are public by default in Twitter. Twitter is indexed with the popular search engine like Google, Bing. Hence every user is accessible

to the public. Twitter communication is primarily through tweets - a word set that can contain up to 140 characters. These tweets are filled with short text data consisting of modern world dictionary. Users can tweet about any topics, such as current news, sports, politics, relationship status and holiday plans. There is no upper bound on the tweet count.

The following conventions enhance our understanding of twitter better for this research.

Twitter doesn't have a concept of Friends. "Follow" concept is followed in twitter, where we just have to follow any person we want. Hence, to map the friends list, we used the intersection list of the followers (people whom the user follows) with followees (people who follow the given user) as friends list.

A tweet starting with RT is called Re-Tweet. This means that the user has re-tweeted someone else's tweet. Most of these tweets are about the general events and happenings. We ignore them, because we are concerned about family information.

The "@" - tag is used to tag a user. @ - followed by a user name is the convention used to tag a user to the message. If a user is tagged to a message like "Happy birthday mother", the probability of that user being a mother is high. Hence, we mine this @ - tag information to look for family details.

2.2 Preliminaries

2.2.1 Twitter API

In order to collect twitter data, we used the Twitter open-source API - twitter4j [5]. twitter4j allows us to collect the publicly available user information such as friends, followers, tweets by providing the user id. This API holds an upper limit

of 3200 tweets per user. Therefore, we could collect only the latest 3200 tweets of any given user. User's who lock their account private cannot be accessed through this API. Hence, in our research we have not considered the private accounts.

2.2.2 Standford NLP Tagger

Parts-of-speech recognition is one of our major modules. We used Stanford NLP tagger [6] is order to achieve the sentence's POS tagging. We use the English Dictionary tagger for the POS tagging. This API tags non-dictionary words with approximation. Hence, this tagger is useful with the short text micro-blogging data. Below given example is the output from this tagger.

Table 2.1. POS Tagger Examples

Base Text	Output Text
happy birthday mom. You are the best	happy_JJ birthday_NN mom_NN ._. You_PRP are_VBP the_DT best_JJS
my wife looked lovely today	my_PRP$ wife_NN looked_VBD lovely_JJ today_RB
happy bday mommy. U r the best	happy_JJ bday_NN mommy_NN ._. U_NN r_VBP the_DT best_JJS
Hate you god!!! boring life and a boring family	Hate_VBP you_PRP god_NN !_.!_.!_. boring_JJ life_NN and_CC a_DT boring_JJ family_NN

2.2.3 UMBC Similarity Tool

For similarity calculation, we have used the tool developed by the research team at UMBC [7]. We use their Phrase similarity [8] and Semantic similarity [9] in our research.

2.2.3.1 Tool Overview

Below observations are some of the characteristics of this tool related to the research.

- Produces better results with smaller case letters.
- Singular/Plural is treated the same.
- Yields better result with URLs removed from the text.
- Extra whitespace characters, does not affect the results.
- Numbers in-between the words are not processed.
- Non-English words reduce the resulting result by multiple folds.

2.2.3.2 Phrase Similarity

For the Phrase similarity, the corpus options available are the Stanford WebBase Corpus and the LDC English Gigawords corpus. We have used the Stanford WebBase Corpus with the Relation similarity since we achieved better results with it. We use Phrase similarity to identify the direct noun phrase match for relations.

2.2.3.3 Semantic Similarity

For the Semantic similarity, we use the GetStsSim [9] API to get the similarity score. This API works based on the Distributional Similarity and LSA Similarity. They use WordNet for boosting their algorithm.

2.3 Related Work

Mining micro-blogs is an upcoming field dominating this era. Rich source of information present from this industry attracts commercial prospectus. Below given are some of the related work focusing on different attributes of OSN's.

2.3.1 Privacy Related Work

Privacy related study in micro-blogs is a major research area because of the huge amount of data present and sensitivity of the data. Most of the research performed in this area follows direct term match, words frequency and pattern observation based information retrieval. Research pertaining to this area focuses on identifying the private attributes of the user as our research does.

For both public profile and private profile users, [10] developed a prediction model to predict their birth year (their age) based on his/her graduation year, reverse lookup, friends-of-friends age. One of the micro-blogging sites - Facebook was used as their data feed. They proposed a privacy change in Facebook,*"When Alice chooses to hide her friends in her limited profile, Facebook should also automatically remove Alice from the friend lists in all her friends limited profiles."* [10]. Facebook currently implements this feature partially It hides the user name for two uncommon friends, whereas it shows the user name between two common friends. As in our algorithm, friends of the user play a significant role in predicting the age of the private profile users, but their research is centered towards extracting the openly available details and features rather than mining patterns.

Another study by [11] is focused on identifying vacation plans, medical condition, and alcohol influential tweets of the users. [11] has also focused on privacy violation in Twitter and its impacts that end-users are not aware of. They have

identified the privacy leaks in twitter because of Re-Tweet feature - where private tweets are exposed as public based on user settings, Google Indexing - with just a screen name, all the tweets about a person could be found without getting access into Twitter and the default twitter settings - all posts are public by default. A part of their research also briefs on family related tweets that breech privacy. Their data-set provides $< 1.6\%$ classification for family related tweets. For predicting the medical condition, their research is pattern based where positive patterns look for keywords like cancer, disease etc. and noise filtering negative patterns dogs, cat etc. to remove noisy data. Their research is term based information mining on Twitter data designed to identify sensitive tweets in helping drink and drive and medical emergency scenarios.

Insightful study on twitter data is discussed in [2]. Their content analysis based research delves deeper into data patterns and privacy patterns in Twitter data. Their research confirms the absence of private attributes like SSN, Phone number , Address across majority of users. Their results also indicate the presence of private information such as location, events, whereabouts in majority of twitter users which could be used illegally by hackers. Their content based approach looks into users data and users tweet whereas in our research we have also considered the friends list in evaluating privacy.

2.3.2 Location Prediction Work

Location prediction is currently the most concentrated upon industry mainly because of the profit that this research offers. Spreading current happenings in a city to the people living in it is a great commercial aspect. Great amount of research is being conducted to predict user's Geo-location based on his/her

micro-blogging information. Twitter Geo-Tag attribute is made use of in all of these research. Unfortunately very less amount of users use geo-tagging feature in Twitter [12] and hence heuristics based research is done in this area along with geo-tag information.

[13] [14] have used an ensemble learning based approach in predicting user location. They take into account tweets, their timezone information, activity information related to timezone, external location knowledge in prediction the user location. They built a content-based heuristic classifier which looks into the count of the places mentioned and visited and the sparsity of the places visited in predicting the home location. Their heuristics assume that the user would talk about his location more than the visiting locations. Though we do not use heuristic based learning, our seed patterns are built using the heuristic observed in twitter data related to family.

Research by [15] predicted the location of the user based on his/her follower/following list and their interaction. They built a decision tree model based on the relationship, closeness, friends location, and the estimated distance between the user and their friends. They then train this decision tree through an Maximum Likelihood Estimator which gives the location of the user with just 21 miles variation at most.

Research based on tweet mining is addressed in [16, 17]. They mine the openly available data, to identify place names, latest events specific to a place, Users profile, their review post related to restaurants, local bars etc in order to predict the user location. They use Bayesian Probabilistic models from words in tweets to predict location of the user.

Geo-Location prediction was also done by [18]. They use purely the knowledge

from the tweet(IP address, login information was not used). They identify the geo-prediction based words from the tweets. Their prediction model predicts user to k locations based on a lattice based smoothening model.

Non tweet based location prediction is studied by [19, 20]. This involves URL crawling, Foursquare check-ins. They use Multi-map API or GPS based information produced by foursquare check-ins to predict the user location.

2.3.3 Commercial Data Mining

Commercial micro-blog Data Mining is also a major focus area. A proposal to predict stock rates based on user tweets is considered in [21]. Their approach uses sentiment analysis using mood tracker tools like OpinionFinder (postive, negative mood prediction) and Google profile of mood states (dimensional mood analysis).

Methods to identify product reviews using Twitter and Amazon data is developed in [22]. They created an emoticon dictionary to analyze smileys and their moods, and an acronym dictionary to address short text issues. Their research focuses on identifying sarcasm in the reviews to help product review better. Other research work related to sentiment analysis is covered in [23, 24].

2.3.4 Similarity Based Mining

Our Similarity based learning was greatly inspired from the research by [25]. Their research deals with identifying similarity between short text data and different similarity measures to evaluate their relation. Their similarity measures include lexical similarity (Purely term matching), Stemming, Probabilistic, and BackOff methodology. They analysis, describes a clear outline of each similarity measures strength and weakness points and scenarios where one measure is useful

than the other.

Another similar research was done by [26] in short text area. Since Cosine Similarity would prove less efficient because of lesser word length, they have used search engine to enhance the short text by added the results retrieved from the engine. They developed a web-based kernel function to calculate the similarity of the enriched short text. Their research helped us to focus on semantic similarity rather than term based similarity.

Chapter 3

The Problem

3.1 Problem Definition

Manually reading each tweet and classifying it as family is an almost impossible task. This research is intended for two purposes.

- Identifying all the tweets related to the family. This includes a minor sub-module to classify the tweets into different family categories.
- Identifying femaily members related to the user under study from the tweets. This module involves predicting user names of the relations found from the tweets.

3.2 Challenges

Dealing with Twitter data imposes lot of restrictions. Some of the challenges are described below.

- Tweet Count

  First major challenge is the tweet count. With the maximum set at 3200, it is difficult to identify all the relationships. The results would be better if all the twitter feeds were made available because many birthday and anniversary tweets might not necessarily belong to the latest 3200 data-set.

- Noisy Data

  It is one of the major concerns in this research. In the absense of noise, this research would have been accomplished by simple keyword matches. Noise data such as "How I met your mother", "Mother Nature", "Mother Teresa" occupy the major portion of the data extracted through keywords. Table 3.1 shows some of the noise tweets found in our data-set. Our method reduces this noise by more than 70%.

Table 3.1. Noise Tweets Examples

Noisy Tweets
stop killing mother nature
love you barney. #HIM yo MOM
christina aguilera and britney spears should start a sister band, they rock together.
OMG! My friends wife looks sexy!!*wink* *wink*
lovey dovey feeling :-). Robin soo suits as Barneys wife.

- Short Text Language

Around 86% of users from the data-set are filled with shortened words and abbreviations. Since short text pattern analysis is not the high-light of this research, we limited our work in this area. Our algorithm uses the words as it is without any processing. This may have led to lesser through-put value. This is one of the focus areas for the future. We have done manual word expansion for certain words related to the family. Words were chosen based on the popularity across twitter.

Some of the most popular Micro-blogging words are given below in table 3.2.

Table 3.2. Shortened Words and Abbrevations

Popular Keywords and Abbrevations
kewl
lol
omg
rofl
lmao
haha
mum
luv
paa
sis
Wassup

- Informal Presentation

  Micro-blog messages do not always follow the English Grammatical rules. Table 3.3 shows few examples of bad contructs. For this research we have not altered the messages and we use the output produced by Parts-Of-Speech tagging as it is.

Table 3.3. Improper English Usage

Tweet Examples
U me friends??@ladygaga
me in love
Me going to Vegas!! Yay!!!!!
Long time no see.
Where u been?
what up buddy?
me and my sis rock! shopping day!!

- Large Volume

  Every user considered for this research holds a minimum of 2500 Tweets. To process each message and to find its similarity measures across various seeds is a time consuming process. Hence, the current process time is a minimum of 3 hours for each user.

- Auto Feeds

  Most of the twitter feeds contain automatic messages posted by Apps such as FourSquare, TwitterStats, Instagram on user's behalf. These feeds are

of minimal use for this research. So we eliminate any URL's along with common automatic contents.

Figure 3.1. Auto Post Tweets

- # - Tags

  Twitter is famous for '#' - Tags, where users categorize their messages named after the '#' symbol. Figure 3.2 shows an example hash message.

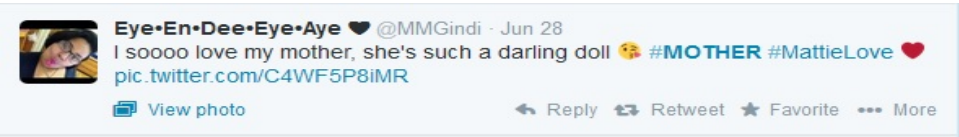

Figure 3.2. Hash Tagged Tweet

  Mining # - Tags may lead to increased results. We have not considered # - Tag information in our work. We truncate the # - tags present in the tweet before processing it.

- Special Characters

  Certain tweets replace English characters with special characters. These special characters were not handled since they might lead to conflicts if

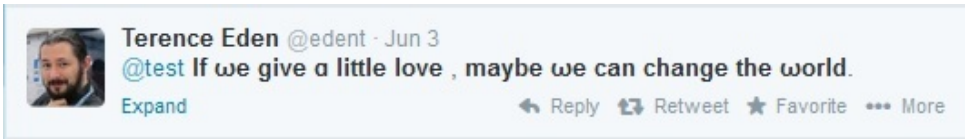

Figure 3.3. Special Character Embeddeds Tweet

intended for math purpose tweets. An example of this tweet, is shown in figure 3.3.

Chapter 4

Algorithm

4.1 Overview

Figure 4.1 gives the overview of our solution. In figure 4.1 inner circle represents the process for each tweet and the outer circle represents the process for each user.

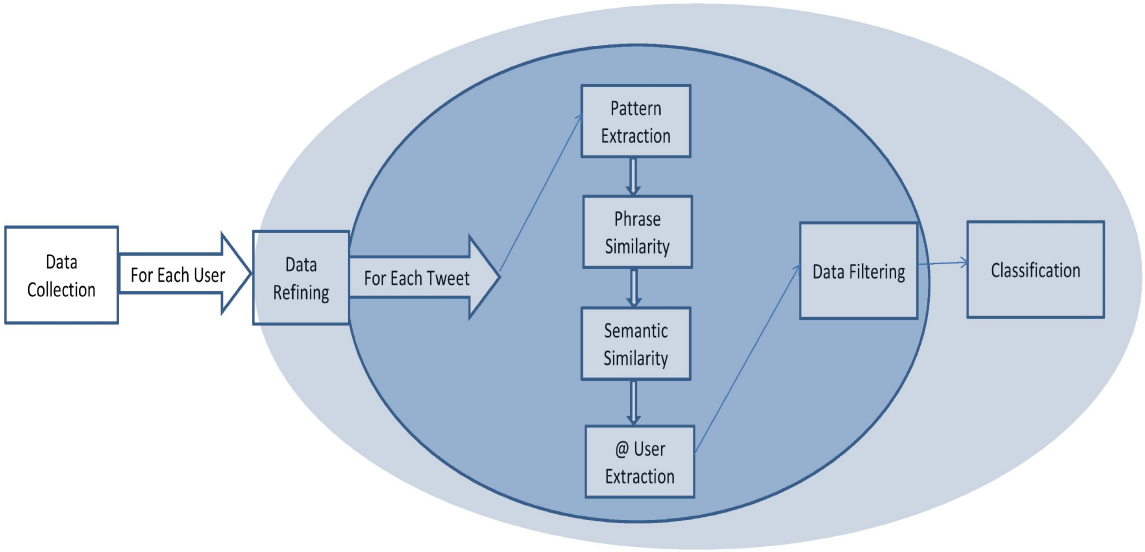

Figure 4.1. Algorithm Overview

The entire implementation of our algorithm is developed using Java. Each tweet is pre-processed to remove all the special characters and other unwanted contents. Our algorithm does not crawl the URLs present in the tweet messages. It does not process image, audio, or video data. We only consider the text and hence discard the rest in the preprocessing phase. Once pre-processed, we run each tweet across different similarity measure to identify its similarity to family relation. The final score after processing through all the phases would be in between 0 and 1. Currently, we consider tweets will score greater than 0.62 as related to family. Finally, we classify each tweet either as related or as unrelated based on the score obtained.

The following sub-sections address each of the components described in this figure.

4.2 Data-set

We collected 150 twitter users' user name, screen name, friends list, tweets and tweets time-stamp. Twitter does not have a special category called friends. Hence, we considered the intersection list of followers and following as friends list. Randomized Algorithm is built to select users with the following criteria,

- Users with more than 1500 followers are omitted as they have more chances of being a celebrity.
- Users with less than 2000 tweets are filtered as it is difficult to predict family attributes on a smaller data-set.
    - 20 users with lesser tweet count were collected for analysis purpose.

- Users with majority of tweets in foreign languages (Anything other than English) are discarded as this algorithm is restricted to handle English language.
  - Twitter holds a language attribute for each user. Some users post content in foreign language though their language attribute is set to English. Since, these users are of minimal use for this research, we read the first 25 tweets and match each word against English dictionary in order to filter better. We consider users, if the ratio of dictionary matches count over the total number of words is greater than 80%. 80% is considered appropriate in order to accommodate short text language.

4.3 Pre-Processing

The raw text extracted from twitter is polished further to suit the algorithm.

- Word Expansion

  Twitter data holds multiple spellings for the same words. Not every spelling would match the dictionary, because of the short text data present. Few examples are given below in Table 4.1 . Hence manual word expansion is done to expand these family related terms to a common word since certain features cannot process incorrect words. Family terms are decided based on manual twitter crawling and finding popular usage.

- Special Characters

  All special characters other than the alphabets and numbers are truncated. Though we do not do any processing with numbers, we still hold it to identify

Table 4.1. Word Expansion Examples

Base Word	Expanded word
mum	mother
gf	girl friend
paa	father
sissy	sister
mommy	mother
bro	brother

patterns related to year, age for future use. All the words are converted to lower case, since the similarity features yield better results with lower case.

- URL Truncation

  Twitter data consists of numerous tweets that has URLs embedded in them. Every tweet from Instagram, FourSquare also has shortened URL leading to their site. Since these URLs are of minimal use to us, we truncate any URLs present in the tweet.

- Stop Words

  Stop word removal is not done for this research since it is important to have words like "my", "our" etc. in predicting relationship.

- # - Tags

  # - tags present in the tweets are truncated. # followed by any list of characters , till a space character is discarded.

- @ - Tags

@ - tags are used for one of the features under prediction model and not the others. Hence, for tweets consisting of @ - tags, a flag is set to indicate its presence and the user name is stored in a data structure for further processing. Finally, the @ - tag followed by the user name is removed from the tweet message.

4.4 Pattern Matching

N-Gram Histogram algorithm is run across the data-set to collect the common patterns containing family terms. All patterns for N values set to since 3,4,5 are collected along with its repeat count. From the patterns extracted, we filter patterns that are most common and that are extensible.

Example

Text 1: *my_PRP$ little_JJ sister_NN* Repeated *48* times

Text 2: *my_PRP$ little_JJ sister_NN @UserName_NN* Repeated *32* times

Pattern: _PRP$ _JJ _NN

Above example shows a pattern that has higher repeat count and is extensible. Hence a list of 40 patterns were collected using this methodology.

Every tweet is processed through the POS tagger and from the output the POS tag is extracted. The extracted tag is matched against the seed tags. If a match is found it returns a positive score else a null value.

Pattern Matching in tweets leads to lot of noise outputs since any phrase could match one of our seed patterns. Hence, Pattern Matching alone provided less useful results since any random text can match the seed patterns. Therefore,

comparatively lesser weightage allocation is given for this phase of task.

4.5 Phrase Similarity

The objective of this phase is to find if the tweet message contains words related to family. Manual verification or hard coded verification is a difficult task since users could use umpteen adjectives to describe a person and countless ways to describe a relation.

To perform the similarity task, we formed a seed set covering all possible relationships with the most frequently used adjectives(Table 4.2). This seed set is currently a smaller fixed set data since each tweet should be compared across all the seed data.

Table 4.2. Common Adjectives

Adjctives
sweet
dear
lovely
little
awesome

We have used the software tool provided by UMBC for calculating the Phrase similarity. This tool takes two phrases and compares the similarity between them just with respect to words present and returns a score in the range of 0 and 1. In order to do this, they use Stanford WebBase Corpus to find possible synonyms of the given word and use that to calculate similarity.

Relative score in the range of 0 and 0.3 is returned based on the value returned from the API. Higher the API return value, higher would be the value returned from this phase.

Table 4.3. Phrase Similarity Examples

Text Compared	Score
Happy Birthday mother *vs.* happy birthday father	0.90173894
Car vs automobile	1.0
Happy anniversary sister *vs.* birthday wishes sister	0.7492
my dear wife *vs.* the love of my life	0.25
grandma is the best *vs.* my life is boring	0.03349926
I love you the most father *vs.* Jesus is great	0.12257728
my sweet little sister *vs.* my handsome young brother	0.37

4.6 Semantic Similarity

Firstly, Semantic similarity will not be performed if the score remains 0 after Pattern Extraction and Phrase similarity. This feature is chosen to identify the Semantic similarity of the tweet. Our aim in this phase is to eliminate the noise data.

Example Tweet: *my dear dog is my best companion*

Pattern Match - Yes

Phrase Similarity - 0.65

Semantic Similarity - 0.33

The above example has a higher Phrase similarity because of the presence of keywords like "dear" and "companion". Although, this text has a higher score it is not a valid family related text. Semantic similarity is useful in such examples, where noise data gets through the other features.

To choose patterns for this model, we used a scaling window methodology. We took a base text "my little sister", and ran the scaling window algorithm on it. This algorithm pattern matches for a scaling window of length 5. It replaces each word and finds the substitution for the given word from the data-set. This is a recurring model and it builds all possibilities along with its count. Though, this algorithm lead to many similar patterns, its sparse nature helped us in forming a varied seed set. Below given is an example of how this algorithm proceeds.

***** little sister*

*my **** sister*

*my little *****

*my **** little sister*

*my little **** sister*

The seed set for this task is formed based on popular messages extracted from the scaling window algorithm. Some of the seed messages are "happy birthday", "anniversary wishes", "congrats", "family time" etc.

Every tweet is compared with seed set data to measure their Semantic similarity. To calculate Semantic similarity, we have made use of the UMBC GetStsSim API [9]. This API takes 2 phrases, and calculates the similarity between them

based on the LSA Similarity and boosted using WordNet. The API returns a value between 0 and 1 as a similarity measure. Similarity score of 0.75 or above indicates a almost perfect match. Similarity score of 0.6 or above indicates a relatively similar texts. Hence, a relative score in a range of 0 to 0.5 is returned from this phase based on the similarity value. Since this method perceives the hidden text, we have given this part highest weightage. Examples of Semantic similarity are given below in table 4.4.

Table 4.4. Semantic Similarity Examples

Base Text	Output Text
happy birthday mother *vs.* birthday wishes mother. you are the best	0.6110
my sweet little sister *vs.* my handsome young brother	0.624
My dear wife *vs.* the love of my life	0.5036315
my sweet sister *vs.* my awesome dog	0.21
long day. i miss you my dear mother. Come back soon *vs.* feeling extremely tired. its a long day	0.38

4.7 @ - Tag Usage

Tweet messages also contain @ - tags, wherein the tweeter tags a person along with the message. This could be a check-in, wishes or just conversation tags.

These @ - tags were processed in the pre-processing stage and a flag indicates the presence of @ - tags. This phase is only performed on the tweets that have @ - tags present in them.

For the user under examination, we maintain a friends list consisting of name and screen name (@ - tag names) of the friends. For @ - tag found on a user tweet, we extract their name from the friends list. If no match found, we discard the screen name.

For the extracted friend name, we verify last name match with the user.

- If last name match is found, we automatically set the @user probability to 100%.
- If not match found, then we keep track of the number of times this @user is found in a data structure along with its relation value.

Finally we use the below equation 4.1 to identify the @ - user relationship along with the probability score of their relation.

$$\boxed{\textbf{\textit{Probability of User - Relation = (N / (U * R))}}} \tag{4.1}$$

- N - Number of times the @ - user is found.
- R - Number of Relations Found, is used to reduce false positives. No user can belong to multiple family categories. Hence this reduces the user probability to minimal in case of multiple relations.
- U - Number of Users under this relation, is used to distribute probability across all users under the same category of relation.
- Note that the relationship classification is done in the classification module.

4.8 Classification

Our final extended module is a regex based classifier. Our classification considers tweets will similarity score above 0.62 and process them through simple regex word patterns. We classify the tweets in to 7 categories listed below in table 4.5.

Table 4.5. Categories

Relation Categories
Parents
Siblings
Children
Ancestor
Relationship
Extended Family
Miscellaneous

Hence the final output is the classified tweets along with the @ - users identified.

Chapter 5

Analysis

We conducted several experiments to evaluate different aspects of our algorithm. We evaluate tweets and @ - tags present. In performing evaluation, we always measure the Precision measure . By Precision, we mean the number of tweets relevant to family from the retrieved tweets. Recall is difficult to measure because of the larger data set and Twitter tweet count constraint.

5.1 Coverage Analysis

5.1.1 Family Tweet Count

This evaluation is just a statistic to show the count of tweets related to the family. We executed the algorithm across 150 users and the total number of tweets considered is greater than 450000. On an average, the output of our algorithm produces 30 tweets per user. A graphical representation of this data is shown below in figure 5.1. Hence, less than 1% of the tweet content is about family members (Table 5.1). This includes the noise produced by our algorithm.

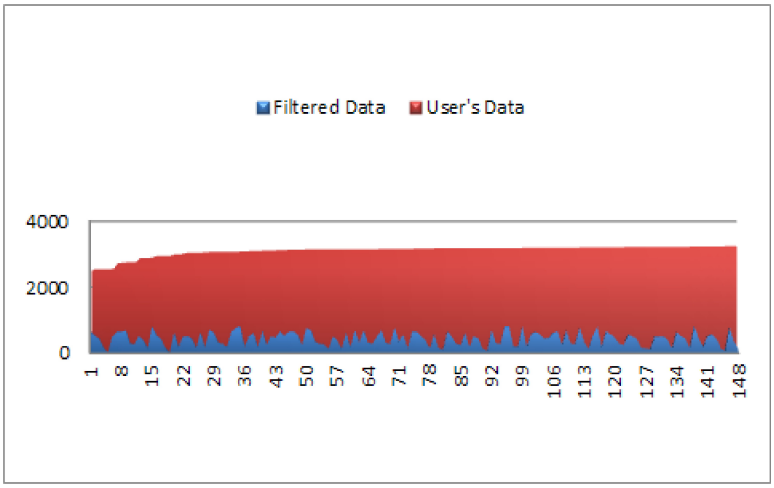

Figure 5.1. Used Tweets vs. Filtered Tweets

Table 5.1. Tweet Count Analysis

Number of users	150
Number of Tweets considered	> 450000
Average number of output tweets	30
Resulting Family tweet Percentage	< 1%

5.1.2 Effect of Tweet Count

Our second evaluation is to show the effect of the tweet count on the output. For this experiment, we used 4 categories of data-set < 1000, 1000-1800, 1801-2400,> 2400 with 20 users in each set. The analysis is as shown below in figure 5.2

5.2 Noise Evaluation

To evaluate the efficiency of our algorithm in reducing noise, we wrote a code snippet, which classifies the data using keywords. Our goal in this experiment

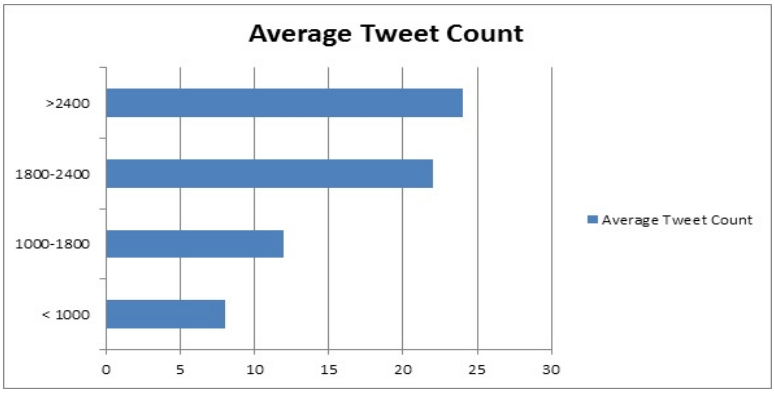

Figure 5.2. Effect Of Tweet Count

was to show the different in performance between our algorithm and the key-word based classifier.

For this experiment, the analysis is shown below in table 5.2.

Table 5.2. Noise Reduction Performance

Total number of users considered	75
Total number of tweets considered	225886
Output Tweets using Keyword Classifier	6121
Output Tweets using our algorithm	2301
Noide Reduction Percentage	62%

This data, may not be absolute since we havent considered the false negative tweets. Even if considered, our performance would still be better than a key-word based classifier.

5.3 Relationship Analysis

We conducted a set of ***manual*** experiments to evaluate the quality of tweets filtered.To do this, we considered the results from 50 users. Our results were restricted to a lesser sample data-set since manual evaluation is a time consuming process. We analyzed the results manually to find the Family Precision value of the tweet messages. Before proceeding to individual analysis, figure 5.3 shows the general tweet pattern from our dataset.

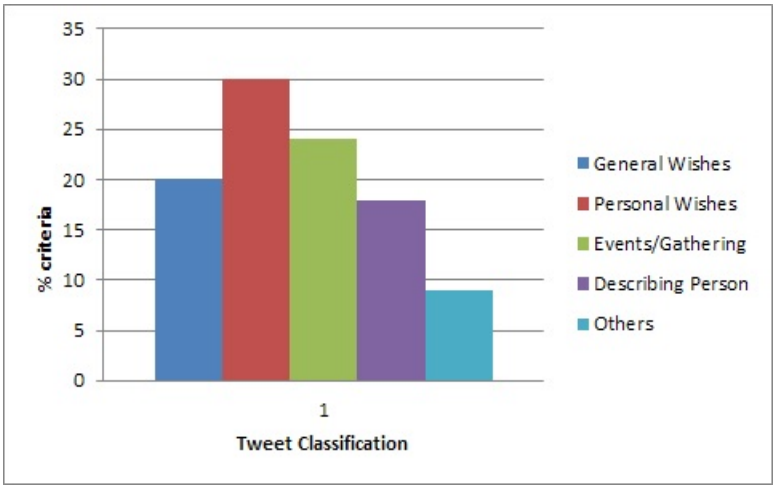

Figure 5.3. Family Tweet Message Distribution

5.3.1 Tweet Relevancy to Family details

For this evaluation, we looked at the filtered tweets only from the family perspective and not very detail specific. Table 5.3 shows examples of the tweets belonging to this category.

@ - tag relationship was not considered for this evaluation. Any tweet, that speaks about a family member or event or occasion is considered a valid output. Table 5.4 shows the evaluation analysis of this phase. Table 5.5 shows few

Table 5.3. Examples of family tweets

Sample Tweet Set
I'm gon be an uncle *smiles* "@Bintah_Adam: I can't imagine my mum having another baby now"
Oh my god my sister is annoying
My granddaddy is dead....
My sister's roommate probably thinks I am the weirdest child
I'm such a daddy's girl.
@BrandonHerreros I was up there with my brother ?

examples of noisy data found.

Table 5.4. Evaluation Of Tweet Relevancy

total number of tweets considered	1346
Total number of tweets found relevant	1110
Percentage of Tweet Relevancy	83%

5.3.2 Tweet Relevancy to Identify Family Person

This experiment is to evaluate the tweet relevancy with respect to identifying the family persons user name. In order to do this, we filtered tweets containing @ - tags in them.

In this evaluation, we only focus upon tweets efficiency and not the @ - tags

Table 5.5. Noise data output

Sample Noise Tweet Set
My friends engaged and no one even knew he was dating someone
When one of my boys tells me he's in love
If your not my girl don't be jealous of my other girls
@Real_Liam_Payne Liam, Its my birthday in a few days! It would so much if you tweeted me a happy birthday and also a follow from you! :(x30
@zaynmalik idk how many times ive wished you a happy birthday, but I hope you get what you wanted and enjoy your day, love you heaps zen xo

efficiency. We want to evaluate *Percentage of tweets that help in identifying relationships* . Good examples and bad examples of identifying relations are given below in table 5.6. The evaluation results are shown below in table 5.7.

This evaluation proved less useful. Hence, we developed a probability based evaluation model.

- Less than 50 % of tweets of the user had @ - tags in them.
- This rate proved lower mainly because of the noise data present, where the person is taking about his surrounding and hence leads to higher Semantic Similarity score.

Table 5.6. @ - Identification Tweets

Good @ -tagged tweets
@DelRioMellisa can you just be my wife already? ?? @PeterThomasRHOA Go Peter Gooooo!!! Congrats on EVERYTHING Brother!!! xoxoxo
Bad @ -tagged tweets
@jossiee_hunnyy awhh my cousin and you look so happy together @RainerMonster Please pray for my little brother. @Lovelyy_Mandaa: Christmas will never be the same without my grandpa. If I could have him back that would be the best Christmas gift. "@LynnLynn47: My brother is so annoying" *Sister

Table 5.7. Evaluation Of @ - tag Tweets

total number of tweets considered	711 ($< 50\%$)
Total number of @ - tags Identified	557
Correctly Identified @ - tags	211
Accuracy of Tweet Relevancy	38%

5.4 Classification Analysis

Though classification is just an added feature for better presentation, we tried to evaluate the precision rate of our classifier. Our classification module is just a hard-coded reg-ex module without any deeper algorithms. For this module, we evaluated the output produced by the classifier. Evaluation is done based on the precision rate of the number of correctly classified tweets. We have 7 categories available for classification as described in table 4.5.

5.4.1 Single Label Classifier

Tweets belonging to single label of output are comparatively easy to classify. Tweets were classified with 76% efficiency under this category. Some of the noise data which lowered the efficiency are given below.

The precision rate is calculated based on miscellaneous classified tweets compared with the total number of tweets. Please note that *tweets containing words like "baby", "girl", "boy" were mapped to miscellaneous section by default, since these terms can branch under multiple labels.*

Noise Data Examples

my girl looked pretty today

Gorgeous Baby! You walked like a princess

boys will be boys! Hating them

5.4.2 Multi-Label Classifier

In our algorithm, every tweet crosses through checks for all the labels. Hence, no special algorithm is written for Multi-Label classification. Analyzing Multi-Label Classifier proved challenging because of the lesser tweets available that

belongs to this category. Table 5.8 shows the analysis of Multi-label classifier.

Table 5.8. Evaluation of Multi-Label Classifier

Total Number of Tweets	210
Tweets Correctly Classified	64
Percentage of Tweets Classified Correctly	$< 35\%$

The noise data present in this category is the "Third Person" noise data.

Noise Data Examples

my brother and his girl are coming over for the weekend

step-dad and his friends daughter are in a relationship:-(.. this world sucks!!

like father like son! yo @JoyceMeyer! U and your father rock!!

5.5 @ - Tag Evaluation

We also evaluated our @-tag module to calculate its efficiency in identifying persons. We evaluated the @- tag on 50 user data set.

5.5.1 Last Name Match Evaluation

Large number of twitter users, use fancy names in their name sections. Few examples are given below in figure 5.4.

For users whose data is under review, we extracted their proper names by doing a Google search with their screen names. Since users tend to use the same screen id across multiple micro-blogs, we were able to identify their names with a little effort of crawling.

Screen Name	Actual Name
DedЯЯick HaяяringtoN	Dedrrick Harrington
זמיר טובי	Toby Zamir
azcat92	Jim Keezell
|| PLEASE MICHAEL ||	Alexis Fraley

Figure 5.4. Skewed Twitter Names

Extending the above method to the friends of the user proved to be difficult, because crawling the search engine for every users 300-1000 friends was time consuming and also provided less fruitful results.

First experiment we performed, is to evaluate the performance of Last name match criteria. Direct Last name match resulted in less than *4 users*. This low result is mainly because of the skewed User names. Table 5.9 shows an example of last name match.

Table 5.9. Last Name Matches

User 1	User 2	Relationship
Ryan Chaffin(rcchaffin)	Tracey Chaffin(tmchaffin)	Mother

5.5.2 Single @ - Tag Evaluation

Secondly, we evaluated @-tags with respect to tweet relationships. Table 5.10 shows the distribution of @ - tags across different categories.

Total number of @ - Tags Considered - 380

Note: There is a difference in numbers when compared to the total number

Table 5.10. Single @ - tag Evaluation

Probability Ratio	Number of Users found	Number of Users Correctly Identified
@ tags with probability score > 50	38	36
@ tags with probability score > 40	41	29
@ tags with probability score > 30	52	40
@ tags with probability score > 20	60	29
@ tags with probability score < 20 and > 0	98	39

of samples considered. This is because, not all @ - tag users are friends because of our friends list assumption. Hence we couldn't find certain @ -tagged users in friends list.

We evaluated the Tweets containing a single @-tag connected to the relationship with a **Accuracy** rate of 58%. The noise here is mainly because of single tweet supporting @ - tag messages. Examples of noisy data are given below.

My kid is in love with the doll @Billy_Heath.

@HeartCapricorn you rock! my wife is so happy! thank god for my beautiful wife.

5.5.3 Multi @ - tag Evaluation

Tweets containing multiple @- tag users proved to be challenging, since they map multiple labels. Pattern matching cannot be done since numerous different

patterns are under study.

Hence, for this research we use the tweet messages of this Multi @ - tagged tweets, but we don't make use of the @ - user names.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

Envisage the scenario where we have to read 3000 messages to identify the details of the user. Eventually, when we finish reading we would forget the initial messages. Our tool helps us to address this issue. The tweet messages dwindle to those related to the family with about 83% precision and with the messages classified. Our algorithm also reduces the noise tweet by 62% approximately when compared with the standard keyword based classifier.

We started our research with focus towards privacy constraint. Our focus still remains the same and this algorithm helps us to identify the privacy breach in OSNs. Our algorithm is a semi-supervised learning algorithm which uses text similarity in identifying patterns. We have used different features to identify patterns from all perspective. The main goal for initiating this research is to reduce the false positive rate. This algorithm does it quite efficiently with a higher precision. Every message produced by this algorithm will be used to identify a friend, family, event or a greetings message. All the features considered in the

algorithm, play a vital role in the tweet prediction and hence try to address the tweet from various perspectives.

This algorithm neither focuses on short text pattern extraction nor on the classification. Our main goal is to filter family related tweets from voluminous data and to identify the family members. The whole algorithm addresses just the filtering perspective. Though @ - tag identification was not exceptionally achieved, we have taken the starting step towards family attribute identification.

6.2 Future Work

The algorithms main intension is to identify the family details. Although it provides positive results it could be improvised further to look for more attributes like hash-tags, friends tweet messages etc. We can also combine the results from multiple microblogs for better results. As a further step, we can also predict the attributes of the family such as their age, location, birthday, isAlive from the tweets extracted. Attribute prediction for certain attributes like birthday, age would be direct, since we already have timestamp for the messages. Hence a single birthday wish message for a family person will reveal their birthdate.

URL crawling may lead to increased results since some URLs lead to pictures of the user (Example- URLs posted by Instagram). Image processing, URL processing will lead to increased results and also to clearer identification of the relative.

Future implementations of this algorithm can have a Client Side protection tool to verify the privacy score of the message to be posted. Another version of a client side protection tool would be to crawl over the users tweet and to predict a privacy score showing the leaks in users account, so that the user can fix every

hole in privacy. Since users are not aware of the privacy breach happening in the internet world, this would be a great awareness tool.

References

[1] Lenhart A., Fox S. Twitter and status updating, Washington DC: Pew Internet and American Life, 2011.

[2] Balachander Krishnamurthy Lee Humphreys, Phillipa Gill. How much is too much? Privacy issues on Twitter. *ICA*, pages 1–29, 2010.

[3] Alessandro Acquisti and Ralph Gross. Predicting social security numbers from public data. *Proceedings of the National Academy of Sciences*, 106(27):10975–10980, 2009.

[4] Haenlein Michael Kaplan Andreas M. The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, 2011.

[5] Yusuke Yamamoto. Twitter Open-Source API. `http://twitter4j.org/en/index.html`, 2007.

[6] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[7] Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Johnathan Weese. UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems, Proc. 2nd Joint Conf. on Lexical and Computational Semantics. Association for Computational Linguistics, 2013.

[8] Lushan Han. Graph Of Relations. `http://swoogle.umbc.edu/SimService/phrase_similarity.html`, 2013.

[9] Lushan Han. Graph Of Relations. `http://swoogle.umbc.edu/StsService/index.html`, 2013.

[10] Ratan Dey,Cong Tang,Keith Ross,Nitesh Saxena. Estimating Age Privacy Leakage in Online Social Networks. *INFOCOM, 2012 Proceedings IEEE*, pages 2836–2840, 2012.

[11] Apu Kapadia Huina Mao, Xin Shuai. Loose Tweets: An Analysis of Privacy Leaks on Twitter. *WPES' 11*, pages 1–12, 2011.

[12] Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. Location inference using microblog messages. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 687–690, New York, NY, USA, 2012. ACM.

[13] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of twitter users, 2012.

[14] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of twitter users. *CoRR*, abs/1403.2345, 2014.

[15] Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. Location prediction in social media based on tie strength. In *Proceedings of the 22Nd ACM*

International Conference on Conference on Information & Knowledge Management, CIKM '13, pages 459–468, New York, NY, USA, 2013. ACM.

[16] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, Eric P. Xing. A latent variable model for geographic lexical variation. 2010.

[17] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. ACM, 2011.

[18] Zhiyuan Cheng, James Caverlee, and Kyu min Lee. You are where you tweet: a content-based approach to geo. ACM, 2010.

[19] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh and Krishna P. Gummadi. Geographical Dissection of the Twitter Network. International Conference on Weblogs and Social Media (ICWSM'12), AAAI, 2012.

[20] Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. Finding your friends and following them to where you are. 5th ACM international conference on Web search and data mining (WSDM), ACM, 2012.

[21] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.

[22] Oren Tsur and Dmitry Davidov. Icwsm - a great catchy name: Semi-supervised recognition of sarcastic sentences in product reviews. In *In International AAAI Conference on Weblogs and Social*, 2010.

[23] Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau. Sentiment analysis of Twitter data. LSM '11 Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, 2011.

[24] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[25] Donald Metzler, Susan Dumais, and Christopher Meek. Similarity measures for short segments of text. In *Proceedings of the 29th European Conference on IR Research*, ECIR'07, pages 16–27, Berlin, Heidelberg, 2007. Springer-Verlag.

[26] Mehran Sahami and Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 377–386, New York, NY, USA, 2006. ACM.

[27] Sweeney, L. Protecting Job Seekers from Identity Theft. *IEEE Internet Computing*, 2006.

[28] Lenhart A., Madden. Teens, privacy and online social networks. 2007.

[29] Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. I read my twitter the next morning and was astonished: a conversational perspective on twitter regrets. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*, pages 3277–3286. ACM, 2013.

[30] Emiliano De Cristofaro, Claudio Soriente, Gene Tsudik, and Andrew Williams. Hummingbird: Privacy at the time of twitter. In *Proceedings of*

the 2012 IEEE Symposium on Security and Privacy, SP '12, pages 285–299, Washington, DC, USA, 2012. IEEE Computer Society.

[31] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 206–214, New York, NY, USA, 1998. ACM.

[32] Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 120–127, New York, NY, USA, 2001. ACM.

[33] Aixin Sun. Short text classification using very few words. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1145–1146, New York, NY, USA, 2012. ACM.

[34] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. pages 667–685, 2010.

[35] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 841–842, New York, NY, USA, 2010. ACM.

[36] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual*

Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, HLT '11, pages 581–586, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.